Ebbs and Flows of Polarization During a Political Campaign

Kellin Pelrine[1,2], Anne Imouza[1,3], Gabrielle Desrosiers-Brisebois[1,3], Zachary Yang[1,2], Sacha Lévy[1,2], Aarash Feizi[1,2], Jiewen Liu[1,2], André Blais[3], Jean-François Godbout[3], and Reihaneh Rabbany[1,2]

[1]Mila
{kellin.pelrine, reihaneh.rabbany}@mila.quebec
[2]School of Computer Science, McGill University
{zachary.yang, sacha.levy, aarash.feizi, jiewen.liu}@mail.mcgill.ca
[3]Département de science politique, Université de Montréal
{anne.imouza, gabrielle.desrosiers-brisebois, andré.blais, jean-francois.godbout}@umontreal.ca

Abstract

This paper uses a joint embedding model, INTERPOLAR, based on graph convolutional networks and RoBERTa, to combine textual and social network analysis and estimate the partisan orientation of users of the micro-blogging sites Twitter and Parler. We focus on the 2020 US presidential election and its aftermath. By combining information on the structure of the social networks, likes, hashtags, re-posts, and the content of messages, INTERPOLAR estimates the partisan orientation of users who were active during the campaign. We also present the novel INTERPOLAR INDEX for estimating the degree of partisan polarization on a daily basis throughout the campaign, grouping users by ideology and analyzing changes in cluster distances. Our estimates are based on over 4.5 million posts and the user interactions taking place within them. We validate our results through both synthetic and real data analyses, ranging from controlled tests of particular aspects of the model to full-scale comparisons with other methods over months of data. We also use several gold standard measures, such as the voting records of Members of Congress, the party affiliation of users, and primary registration records. Preliminary findings indicate that polarization increased after the 2020 election, with important shifts around the Capitol Hill riot.

Acknowledgement

Paper presented at the *American Political Science Association* Meeting, October 3, 2021. This project is supported by CIFAR through a CIFAR AI Catalyst Grant: Being Politic Smart in the Age of Misinformation. The first author is supported by a fellowship from IVADO.

1 Introduction

The US 2020 election has been one of the most divisive in recent American history. The party system has become extremely polarized over the last thirty years, but this has reached unprecedented heights as Democrats and Republicans now strongly mistrust each other [Gelman et al., 2008, Iyengar et al., 2012, McCarty et al., 2016]. Our goal in this study is to document the ebb and flow of this partisan polarization by analyzing the social media activities that surrounded the 2020 presidential election campaign.

There is a common understanding that social media platforms played an important role in increasing polarization around the election. With much of the campaign activities moving online because of the pandemic, these platforms provide rich data sources into the pulse of society, the public and elite alike. Motivated by this and inspired by the prior works of Barberá et al. [2015], Barberá [2015], Rheault and Cochrane [2020], and others, we examine the activities of users including the presidential candidates, Members of Congress and their followers, on the

micro-blogging site Twitter, as well as the more conservative social media platform Parler.

More specifically, we collected around 350 million tweets and 6.5 million Parler posts, which reflects the activity of the mass public and the politicians (elite). Our full data include over 20 million individuals on Twitter, over 550k on Parler, as well as 540 members of the elite on Twitter. Our data spans over several months, and we focus on the period just before the November 3^{rd} election to the end of January 2021. We record with whom social media users interact: quotes, re-posts and mentions. We also collect their hashtags usage as a proxy for their word choices as well as the actual text of their messages to fully monitor the language they used. We distinguish and study partisanship at the elite and at the mass levels to compare and validate our method. Figure 1 provides an overview of our methodology (explained in Section 3) to map the user's activity to polarization based on posts streamed from Twitter. The same method can be applied to posts from Parler, although the terminology is different, e.g. retweet v.s. echo.

Our first task is to develop a measure of partisan polarization. We start at the elite level on Twitter. The set of political actors is defined as the Members of Congress and the two presidential and vice presidential candidates. We examine their 'conversations,' that is the words or hashtags they use on social media, as well as whom they refer to. From these, we estimate the location of each actor on an underlying dimension representing partisan conflict. Using our joint embedding model INTERPOLAR, we predict the party affiliation of politicians with very good accuracy. Our embeddings also contain information on the polarization of their voting records in Congress (as measured by DW-NOMINATE scores). These results indicate that our model provides a valid representation of elite polarization.

We follow a similar approach at the mass level. We retrieve the Twitter posts of users related to the 2020 US election. We further identify liberal and conservative users from their profile by using a keyword filter and a language model classifier to construct a profile-based label of party identification. We validate this model on users labeled by experts, and then use its predictions to train INTERPOLAR, which learns from users' posts rather than their profile, by following the same procedure as for the politicians. We show that this model is accurate at predicting the partisan orientation of non-elite users. We also validate our measure by using primary election voting records. Here again the model delivers solid predictive performance. This confirms that INTERPOLAR provides a reliable tool to measure mass polarization.

Besides using real world data, we also conduct controlled synthetic experiments to verify if our model is behaving correctly. We find that it works as expected, and motivate an important modification to resolve a potential source of measurement error due to variable party imbalance (when users of one party are posting more than another).

Our second task is to describe how partisan polarization fluctuates over time and to determine if events surrounding the election contributed to increase partisan polarization. Here, we are interested in identifying specific campaign and post-election events that may have increased or reduced polarization.

The paper offers three main contributions. First, we present the new joint embedding method INTERPOLAR and use it to predict the party affiliation of social media users. We validate these predictions by multiple experiments with real data. Second, we use INTERPOLAR to measure partisan polarization over time and at scale, and validate it with a combination of real and synthetic experiments. And finally, using the above, we analyze changes in polarization around key events related to the 2020 US election and its aftermath, such as the January 6^{th} Capitol Hill riot.

Our paper is organized as follows. In the next section, we present a definition of partisan polarization on social media and review related work. The following section introduces the data and method used in the three main experiments of our analysis. The fourth section presents the results at the elite and mass levels, as well as our dynamic model results analyzing polarization over four months surrounding the 2020 US election. In the last two sections we discuss these results and conclude.

2 Background

Partisan polarization has traditionally been measured by either looking at the difference between the policy positions of party members (spatial polarization) or by focusing on how much they dislike the other

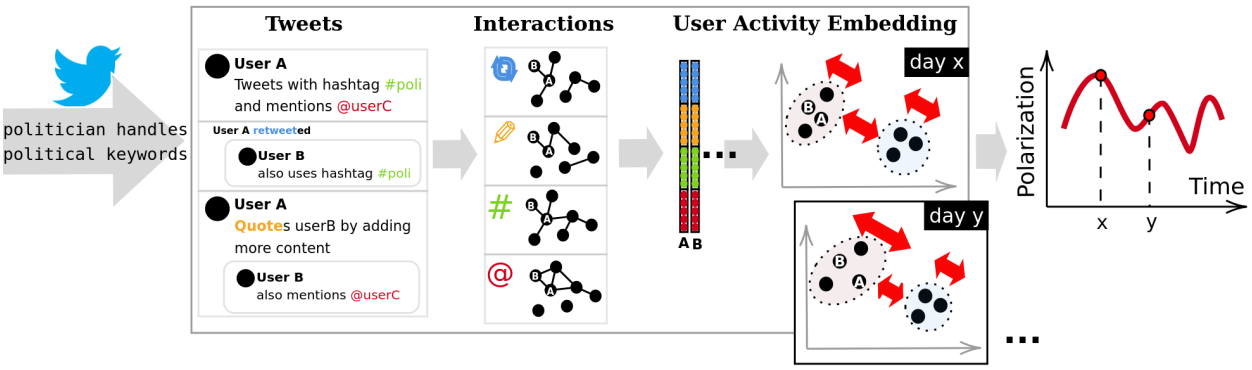

Figure 1: Mapping users' activity on Twitter to create the INTERPOLAR polarization index.

party (affective polarization). In spatial terms, polarization implies a movement away from the center towards the extremes, where distances are measured on a scale representing the left and the right of the ideological spectrum [Fiorina and Abrams, 2008]. In affective terms, partisan polarization focuses on citizens' emotions; it corresponds to the intensity of negative/positive feelings towards politicians or political parties [Gidron et al., 2020].

By using both measures, scholars have found that partisan polarization has increased over the last thirty years in the US, both at the elite level and in the mass public [e.g., Iyengar et al., 2012, McCarty et al., 2016]. Elite polarization has primarily been confirmed over time by looking at trends in the behavior of Members of Congress through the spatial analysis of legislative voting records. Different scaling techniques of roll-call votes, like DW-NOMINATE scores [Poole and Rosenthal, 2007], have been used to estimate the ideological locations of Democrat and Republican representatives in a multidimensional policy space. Poole and Rosenthal [2007] have demonstrated that Representatives and Senators shifted their position to become increasingly distant from one another in recent years, which they take to be a sign of growing partisan polarization. Mass polarization, on the other hand, has been shown to exist in both affective and spatial terms through the analysis of public opinion survey data, either by comparing the policy preferences of Democrat and Republican party identifiers [Layman et al., 2006, Fiorina and Abrams, 2008] or by contrasting citizens' affective ratings of the different parties [Iyengar and Krupenkin, 2018, Hetherington and Rudolph, 2015].

Whether one looks at polarization in spatial or affective terms, measuring partisan divisions at the elite or mass level can either be done cross-sectionally, by looking at a specific point in time, or through time, by using multiple data points or panel data. As DiMaggio et al. [1996] and Fiorina and Abrams [2008] argue, it is particularly important to detect changes in polarization over time, as opposed to taking a snapshot of the distribution of policy preferences in a survey, since a dynamic measure of partisan divisions can help us understand what type of events produces changes in public opinion.

The structure of data used in this paper allows us to capture both the level and the change in partisan polarization before, during, and after the 2020 presidential campaign. By using social media information from the micro-blogging sites Twitter and Parler, we measure partisan polarization at the elite and mass levels by identifying all of the Members of Congress and presidential candidates, and by looking at the behavior of users who are interested in US politics, which we assume to be a subset of partisans in the American population.

Our definition of partisan polarization combines both spatial and affective dimensions. We assume that partisanship is encoded in the choice of whom to follow/re-post/mention and similar social media interactions. These patterns of behavior relate to "homophilic" conversations between like minded individuals, as opposed to "heterophilic" conversations, which refer to the occasional exchanges between people who have weaker social ties [Yarchi et al., 2020, Barberá, 2015]. Our approach assumes that

partisanship is related to word choices, that is which hashtags are used, which topics are discussed, and the vocabulary included in social media messages. The assumptions here are similar to the ones related to social interactions: like minded individuals tend to use the same type of vocabulary and the differences in word choices between Republicans (conservatives) and Democrats (liberals) should reflect partisan (ideological) divisions [Slapin and Proksch, 2008, Diermeier et al., 2012, Gentzkow et al., 2019]. Like Rheault and Cochrane [2020], our analysis is based on a neural network framework to combine both the information about the word choice and the network structure into a single model to estimate the ideological placement of social media users. We thus define:

> **Interactive polarization**, as the difference between the overall vocabulary and interactions observed within or across partisan or ideological groups. That is, the more partisans differ in their social media interactions or word choices, the greater the polarization. Conversely, the more they share or use similar language, the weaker the polarization.

This definition is in line with other approaches to estimate partisanship from text contained in online messages [Green et al., 2020, Grinberg et al., 2019, Gruzd and Roy, 2014, Yarchi et al., 2020]. It also naturally applies between two (or more) groups; for example between Democrats and Republicans or liberals and conservatives. It can be considered within a group as well, to identify people who are more extreme compared to the rest of the group. It is important to distinguish between partisan and ideological polarization. In our study, we examine elements of both, but mainly ideological polarization. This is primarily a result of various users we labeled as liberal or conservative, and then used as building blocks of our models and analyses. So our main experiments focus more on measuring the distance between liberals and conservatives. That said, our model can also work effectively with party, as shown in analysis of predicting user party according to voter records.

Below, we describe in greater details how we estimate individual polarization scores by examining text but also network interactions in a joint embedding model. However, we first review in the next section related work which estimates partisan polarization from social media data.

2.1 Related Work

Scholars have adopted three broad classes of models to measure partisan polarization from social network content. The first is based on the words used by users in their posts on social media. Here, researchers usually rely on a set of specific keywords in dictionaries to identify political messages and code their political leanings [Gruzd and Roy, 2014, Grinberg et al., 2019]. The political leanings can also be inferred from the text used in social media posts by using word embeddings [Conover et al., 2011, Yang et al., 2017]. This type of analysis is useful for detecting the main issues raised on Twitter (what people talk about), as well as the degree of partisanship contained in these messages (how they talk about it) [Green et al., 2020].[1]

The second approach relies on the information provided by the network of users, who they follow, and who follow them in return [Conover et al., 2011]. This method is by far the most popular to infer the ideological leanings of social network users. Barberá [2015] offers the best example of this type of analysis by estimating the left and right position of Twitter users through an item-response model, where the decision to follow a particular user is a function of ideology, the popularity of an account, and political interest. This model is then able to locate relevant ideological clusters on Twitter and confirms that users are more likely to interact with liked minded

[1] Several studies [Gruzd and Roy, 2014, Yang et al., 2020, Grinberg et al., 2019, Yang et al., 2017] have attempted to measure the ideological orientation of Twitter users by looking at the specific textual content of their messages. Some of these studies [Gruzd and Roy, 2014, Grinberg et al., 2019] have relied on sets of specific keywords to infer political tweets and their sentiment/political leaning. Similarly, Yang et al. [2017] and Yang et al. [2020] have relied on semantic representation of hashtags using the "word2vec embedding" in order to measure the average difference between or within specific tweets aggregated by groups to infer users' ideological alignment on a left-right scale. Finally, one study [Badawy et al., 2018] has determined the political ideology of Twitter users based on the political leaning of the media outlets they shared on their profiles.

individuals.[2]

Finally, a third group of models focuses more on the dynamic aspects of polarization in social networks over time by using either one of the two approaches described above. For example, Barberá et al. [2015] constructed a daily index of polarization by relying on the network of users to demonstrate that certain events, like the Newtown Shooting in 2012, increased ideological conflict between liberals and conservatives on Twitter. On the other hand, authors like Green et al. [2020] used the text features of social media messages to build a dynamic measure of polarization over time. In this last study, the authors trained a random forest machine learning algorithm to measure the level of elite polarization on Twitter during the 116^{th} Congress. Their results confirm that there was a surge in the level of polarization on COVID-19 related tweets, with Republicans becoming more distinctive in their behavior than Democrats in the early months of 2020.[3]

In this study, we propose INTERPOLAR to combine all three of these approaches into a single model using a joint embedding framework. Our goal is to estimate the underlying ideology of social media users by looking at the content of their messages and their networks, but also to determine if party polarization has fluctuated over time during the most recent US presidential election.

3 Method

In this section, we describe how we collected the data on the social networks Twitter and Parler. We also explain how we identified elite and non-elite users to construct measures of partisanship. Finally, we discuss the methodology to estimate the underlying ideological orientation of each user, which serve as a basis to develop our dynamic measure of polarization during and after the 2020 election campaign.

3.1 Data Collection

We curated five datasets, summarized in Table 1. In this table, the users are the authors of the posts, while the nodes represent users in our interaction graphs (see Section 3.2.1)—the authors plus users that are referenced within the posts. The hashtag, mention, retweet (or re-post), and quote columns, indicate the number of edges connecting the nodes.

Twitter We collected all tweets, retweets and replies from 995 elite accounts linked to the public and personal Twitter accounts of the US representatives (433), senators (99), as well as vice presidential and presidential candidates (8) using Twitter's Search API.[4] We call this the Politicians dataset.

We also collected around 1% of real-time tweets using Twitter's streaming API, that included one of the following US election related keywords: [JoeBiden, DonaldTrump, Biden, Trump, vote, election, 2020Elections, Elections2020, PresidentElectJoe, MAGA, BidenHarris2020, Election2020]. This constitutes the Election dataset with approximately 350 million tweets and 20 million users. From these, we sampled 20 thousand users, which is the mass Public V1 dataset.

Some days in the Public V1 dataset are missing due to interruptions in the collection pipeline: October 28^{th}, November 17^{th} and 24^{th}, and December 1^{st}, 12^{th}, 13^{th}, 22^{nd}, and 23^{rd}. There are also two days that are partially missing, December 2^{nd} and 9^{th}.

In order to fill in those gaps and extend the time period we cover, we used the Twitter Academic API to retroactively collect all tweets from the users in Public V1. In the period between our original data collection and this retroactive collection done in Summer 2021, we found 5,091 users accounts had

[2] Another closely related approach relies on defining general ideal points of moderate Democrat and Republican Senators by using roll call data [Chen, 2015]. Other studies have looked at the ideological distance between users by observing patterns of interaction among party followers in Europe [Bright, 2017, Gaisbauer et al., 2021].

[3] Other studies like Yardi and Boyd [2010] study the issues of gun violence and abortion on Twitter over a period two months and state that homophily may impact polarized discussions online. Badawy et al. [2018] estimate a dynamic measure by asserting a political leaning to each user regarding the media outlets they share over a period of two months before the 2016 US Presidential Election. Garimella and Weber [2017] also investigate changes in political polarization on Twitter between 2009 and 2016 by estimating the ideology of users from the type of politicians and media they follow. Their results confirm that polarization has increased over time.

[4] Some Members of Congress have more than one social medial account (e.g., one personal and one official account). In this case, we collected information for all of the relevant accounts.

Dataset	Posts Collected					Interaction Graphs			
	Source	Start	End	Posts	Users	Hashtag	Mention	Retweet	Quote
Politicians	Twitter	2020-08-01	2021-01-17	156,562	995	162,121	212,074	83,920	54,630
Public V1	Twitter	2020-10-26	2020-12-31	2,871,050	20,008	255,895	2,510,444	1,621,304	612,827
Public V2	Twitter	2020-10-01	2020-01-31	4,599,125	20,008	961,112	6,319,784	3,943,346	861,905
Election	Twitter	2020-10-26	2021-01-04	348,671,076	20,533,417	-	-	-	-
Parler	Parler	2020-10-25	2021-01-08	6,546,658	566,486	-	-	-	-

Table 1: Statistics on the collected datasets and the interaction graphs extracted from this data.

been terminated or otherwise become inaccessible. This can occur if users are suspended, or choose to deactivate their account, or otherwise make their profile private.

According to our profile labels (see section below on classification), the majority of the missing users were Republican, 3811 to 1280. We hypothesize that many of these users left around the January 6^{th} capitol attack, either voluntarily or during the wave of suspensions after the attack that included the account of Donald Trump.

Because we cannot retroactively collect tweets for all the users directly, we combine the newly collected tweets with the previous data to get the Public V2 dataset.

Parler We parsed all posts provided by the Distributed Denial of Secrets[5] and WayBack Machine[6]. Posts parsed[7] have an estimated creation date since the data provided contain relative timestamps such as "1 day ago" or "1 week ago." Parler posts (or Parleys) can contain hashtags (#) and re-posted content (echo).

DW-NOMINATE In order to get an exogenous measure of partisan ideology for our elite group of users, we use Poole and Rosenthal [2007]'s DW-NOMINATE scores for House Representatives and Senators who served in the 116^{th} Congress.[8] These scores are obtained from the roll call votes of Members of Congress through a multidimensional scaling procedure. The projected first dimension has been shown to represent the ideological conflict opposing the left and the right—from the most extreme to the most moderate positions. Each member is aligned on this continuum, depending on how liberal or conservative their voting record is. These scores were then matched to the Politicians profiles present on Twitter.

Primary Voter Registration Data Following the previous work of Barberá [2015], we match users found in our Twitter database with the primary voter registration records from Ohio, New York, Florida, Arkansas, the District of Columbia and North Carolina. In the voter registrations, we obtain the party affiliation of unique users in each state by md5-hashing their names and county as the key. From our Election database, we extract all users that provide a location in their profile and run that location through Open Street Map[9] and ArcGIS's API.[10] If both APIs return a latitude and longitude that is within 1 of each other, we trust this user has been properly geolocated. We further filter down users that belong to the specific state and remove those whose county could not be retrieved. Finally, we match the most recent voter party affiliation records from the registration data to the unique Twitter users that match county and either first name last name or first middle last name. We normalize the user's name on Twitter to remove emojis.

After matching, we remove users not affiliated with one of the two major parties and users whose name matches more than one record per county. In experiments so far, we focus on New York and Ohio. Basic statistics are shown in Table 2.

3.1.1 Classification

In order to train and evaluate our models, we classify a sample of users according to their party affiliation

[5] https://ddosecrets.com/wiki/Parler

[6] https://web.archive.org/web/*/https://parler.com

[7] https://github.com/RSTZZZ/parler_parser

[8] https://voteview.com

[9] https://www.openstreetmap.org/

[10] https://developers.arcgis.com/python/

State	Democrat	Republican	Total
New York	4843	1631	6474
Ohio	320	193	513

Table 2: Number of users in the Election dataset matched to their voter records for two states.

and ideology based on the description they provide on their user profile on Twitter and Parler.

First, for each dataset, we classify users as "conservative", "liberal" or "unknown" based on identifiers in the description. For "conservative," we use: [conservative, gop, republican, trump]. For "liberal," we use: [liberal, progressive, democrat, biden]. We label users as "conservative" ("liberal") if the description contains at least one of the conservative (liberal) identifiers and does not include any of the liberal (conservative) identifiers. The rest of the users remain as "unknown." Note here that we combine concepts related to both the ideology and the partisanship to label liberal and conservative users.

This is a "weak" classification because user keywords may not match their actual party affiliation or ideology. For example, instead of a president name indicating support, they could say "I hate Trump" or "I hate Biden." In order to validate the overall performance of these labels, we asked two expert coders to classify, on the basis of the very same information (that is, the description provided in the user profile) 60 users from the politicians dataset, 1000 general public Twitter users from each party, 200 conservative Parler users, and 500 liberal Parler users. This "strong" classification either confirms the weak labels, or indicates the presence of a coding error. Note that while in most cases an incorrect weak liberal label indicates that the user is in fact a conservative (or vice versa), a small number of these users can also be independent or apolitical. After comparing the weak with the "strong" labels, we found that users in the Politicians dataset are generally more politically involved and hence the simple keyword search is very accurate. However, for other users, the accuracy was lower, with only around 70% of the weak labels matching the strong labels.

Therefore, we used the strong labels to train a classifier to generate more accurate labels. We randomly split the strong-labeled data into a 75% training set with Twitter and Parler combined, and a separate 25% test set for each platform. With this data we fine-tuned a roberta-large [Liu et al., 2019] model to predict the party each user is closest to from their profile description.[11] We report the results in Table 3.

Dataset	Counts		Accuracy	
	Cons.	Lib.	Cons.	Lib.
Politicians	1,174	1,068	97.7%	96.8%
Election	183,207	176,271	87.0%	90.5%
Parler	31,966	808	93.1%	82.9%

Table 3: Number of users with explicit party/ideological keywords in their profile description (on the left). Accuracy of our profile label classifier based on a manually labeled sample (on the right).

These results show that the classifier provides a reasonably accurate classification of ideological labels. These profile labels are still imperfect, but they are sufficiently accurate for use in training our model. We note that the classifier is binary, liberal or conservative—it cannot classify a user as moderate or independent. In situations where one of those labels would be more appropriate, the classifier will nonetheless say whether it thinks the user is closer to being a liberal or a conservative. We also note that there are far more conservative than liberal users on Parler, matching the platform's reputation for being almost exclusively favored by conservatives.

3.2 Measuring User Activity

An overview of our approach is shown in Figure 2. We collect each user's posts and profile descriptions. From their profiles, we construct profile labels. From their posts, we extract user interactions. These generate a series of graphs which are then integrated into different deep graph models to produce user activity embeddings. Meanwhile, we also integrate the posts into a deep language model to get user activity embeddings based on the text. We use these embeddings to predict party affiliation and measure partisan polarization.

[11] RoBERTa is a pretrained language model; the large version we use has 355 million parameters. It is based on the transformer [Vaswani et al., 2017] and the BERT architecture [Devlin et al., 2018], with modifications designed to improve the training process.

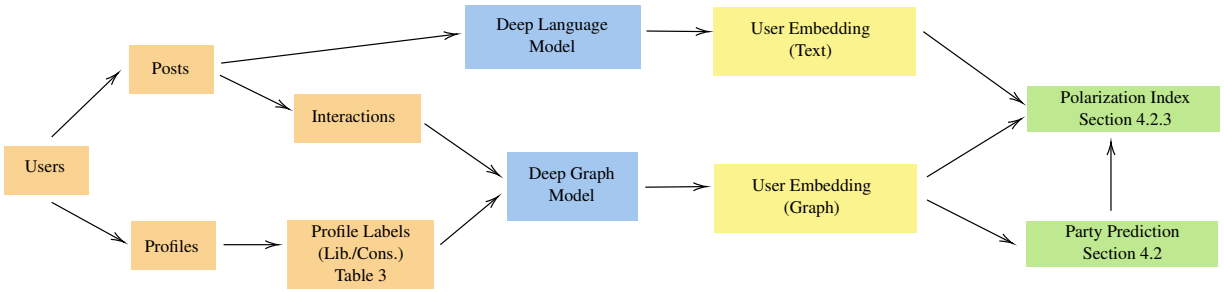

Figure 2: Leveraging users with explicit profile information for mass ideology prediction.

3.2.1 Constructing Interaction Graphs

The next step in assembling our data is to construct interaction graphs—i.e., graphs where the nodes are users and edges represent interactions between them. There are two types of interactions that can link people together:

- Direct links. For example, person A mentions or re-post person B.
- Indirect links. For example, persons A and B both mention the same person C, even though A and B may not mention each other directly.

Direct links are intuitive, however, we find empirically that indirect links are more useful for grouping similar and separating dissimilar partisans. Therefore, we construct graphs with edges based on indirect links between user nodes.

We construct one graph each for hashtags, mentions, retweets, and quotes on Twitter (note: the Parler analysis will be included in a subsequent version of this paper). These interactions can be collected directly from tweet data from the Twitter API. Previous works such as Barberá [2015] heavily used follower networks, but these require separate scraping that can be challenging on a large scale.

In order to get accurate predictions and measurements, some amount of user activity is needed—if a user is not connected to anyone else in the network, our model cannot give a meaningful prediction. Therefore, for a user to enter our interaction graphs we require at least 10 edges connecting them to other users (for example, they use a hashtag which is also used by 10 other people). Second, we filter users who appear in all four interaction graphs. This filter is applied exclusively on the output side—the other users still appear in the training set, but they are not used to evaluate the results.

Most of the politicians are quite active, so we retain 724 accounts of members of congress. There is much more variation in the mass public, but we retain approximately between 50-250 strong-labeled users every day, which we use for evaluating party prediction. When measuring polarization over time, we do not require strong labels, which gives approximately 400-1000 users per day.

3.2.2 Generating User Embeddings

We show the INTERPOLAR modeling process in Figure 3.

We first estimate a partisan position with the text of each user's posts. We use a roberta-large language model (LM). In this case, we want an embedding for each user rather than an immediate prediction of their party or ideology, so we use the pretrained model directly, without fine-tuning on any prediction task. We embed all the tweets, then average the embeddings per user to get a single embedding for each user.

Next, we use the interaction graphs. We do this with a semi-supervised graph convolutional network (GCN) [Kipf and Welling, 2017]. Our model has two terms in the loss function used to train the model, added together with equal weight: one for link prediction, and one for user label prediction. For link prediction, we construct a randomly connected negative graph.[12] The model learns from this in an unsupervised way by trying to predict which links are from the real graph and which are from the negative

[12] An example of this approach from the Deep Graph Library: https://docs.dgl.ai/en/0.6.x/guide/training-link.html

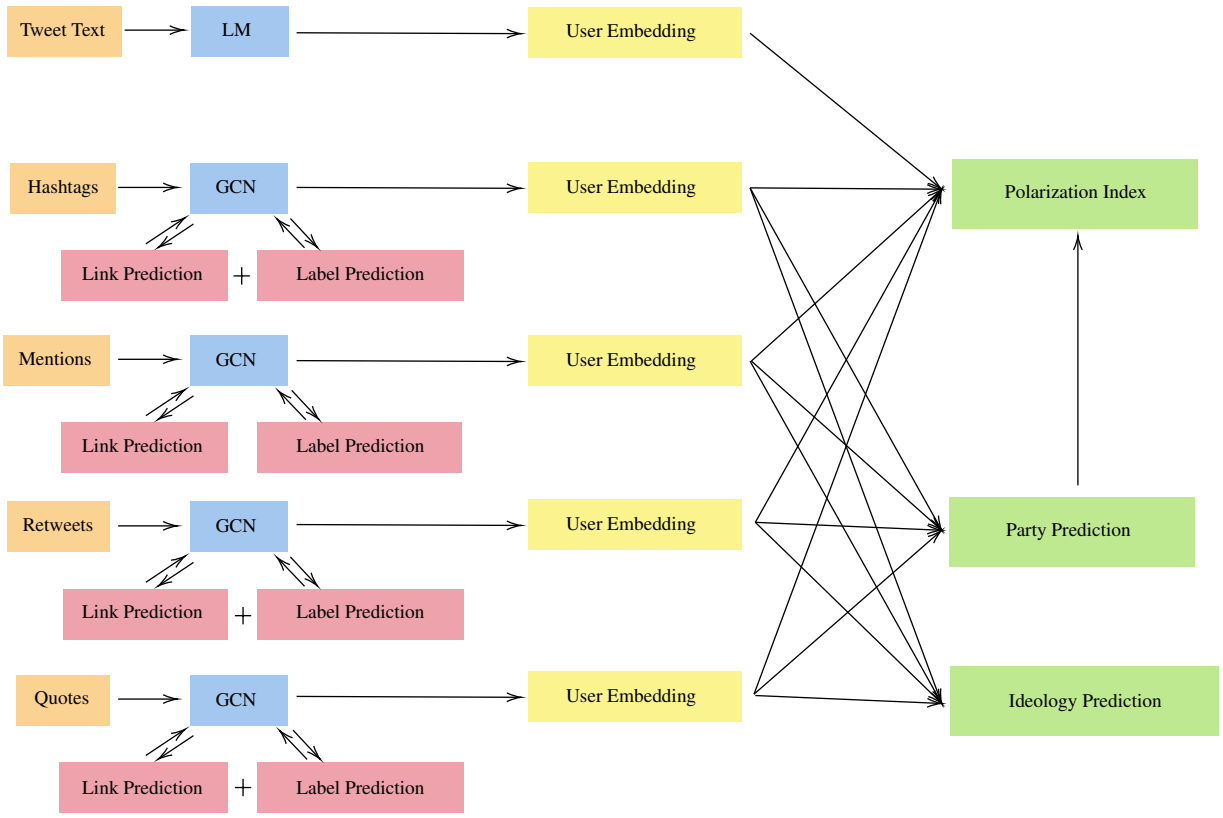

Figure 3: An overview of INTERPOLAR, which combines different modalities of data into ideology and party predictions as well as polarization measurements.

one, using a cross entropy loss.

With regards to the supervised user label prediction part, we use the profile labels (i.e., liberal or conservative) where available for non-politician users, and the true labels for politician users (which essentially match the profile labels, since they are very accurate on this data, as shown in Table 3). The model learns from these labels with another cross entropy loss. Labels in our various test sets, whether profile or otherwise, are withheld during training.

From this model, we get one embedding per user and per interaction type. We can combine these in various ways to predict ideology, party, and measure polarization, as discussed in the following section.

GCNs typically use node features. We initially planned to use the content of the text as discussed above. However, we found empirically that it did not seem to provide any clear benefits. This matches the findings of Xiao et al. [2020]. We simply use a random vector as node features. However, while the text representation does not appear to improve link or party prediction beyond the interaction graphs themselves, or at least not that the GCN is extracting effectively, it likely still contains useful information on user polarization. Therefore, rather than as node features, we use the text embeddings directly as a fifth type of representation.

We train our GCNs for 1000 epochs, chosen empirically with validation data. We use the Adam optimizer [Kingma and Ba, 2017]. All models are trained on an RTX8000 GPU. For each GCN, we produce a 100-dimensional embedding, while roberta-large produces a 1024-dimensional embedding.

3.3 Measuring Ideology, Partisanship, and Polarization

3.3.1 Congressional Polarization

To predict the party of Members of Congress, we concatenate the user activity embedding and pass them to a random forest classifier model, imple-

mented with default settings through scikit-learn [Pedregosa et al., 2011]. We preserve the same train-test split as the GCN.

We predict the ideology of members (i.e., their DW-NOMINATE scores) similarly, with a random forest regression model this time. We find empirically that it is important to include in this model the politicians' parties as an input. This parallels Barberá [2015], who gives party label in the initialization stage of his analysis.

We observed that AutoGluon [Klein et al., 2020] can often produce slightly stronger performance than random forests. However, it is much slower, which can be troublesome: in one of our analysis, we do 100 runs to reduce random variance in the final measurements. Therefore, we have decided to use the efficient random forest approach and we plan to investigate AutoGluon in future work.

3.3.2 General Public Classification

For the general public, we follow the same modeling process as for predicting politician party. For users matched with voting records, we combine their data, one state at a time, with the Public dataset. We use the profile labels on users in that dataset to learn party predictions for the matched users. Separately, we also train on the profile labeled users to predict whether someone is conservative or liberal, and evaluate on the strong-labeled users.

In order to measure polarization over time, besides the overall predictions, we produce predictions for each day individually. This is done by filtering our data down to a single day, running the pipeline on that specific time point and saving the results, then proceeding to the next day.

3.3.3 Public Polarization Over Time

Given the daily user activity embeddings and predicted conservative or liberal label for each user from the previous section, we can measure partisan polarization by adapting a cluster quality metric. Intuitively, the better separated the two clusters are (corresponding to the partisan division between Democrats and Republicans), the more polarized these positions are.

For quality metric, we start with C-index, which compares the dispersion of clusters of data relative to the total dispersion found in the dataset [Hubert and Levin, 1976]. C-index is one of the best performing criteria used for the validation of clustering results [Rabbany et al., 2012, Vendramin et al., 2010]. More formally, it is computed as:

$$C = 1 - \frac{S_{max} - S_w}{S_{max} - S_{min}} \tag{1}$$

Here, S_w is the sum of within-cluster Euclidean distance measurements, which we assume to be linked to spatial polarization [Poole and Rosenthal, 2007]. S_{min} is the sum of the smallest distances between points. S_{max} is the sum of the largest distances between points. A higher C-index corresponds to more concentrated data.

To convert this into a measure of polarization, we first take $1 - C$ in order to get a more intuitive measure that should be high (resp. low) when polarization is high (low). We next apply upsampling to balance the classes. This is critical to avoid the measure varying purely due to variations in the imbalance of number of points in each cluster. It is empirically motivated and discussed further in Section 4.1. We call the result of these steps the INTERPOLAR INDEX.

Next, we compute the INTERPOLAR INDEX for each type of user activity embeddings individually. This gives five series, one for each of the interaction types (hashtag, mention, retweet, and quote), and one for the text. Note the relative scale of each is hard or impossible to interpret. For example, we see that the average INTERPOLAR INDEX for mentions is high compared to text, but this does not necessarily mean mentions are more polarized than text, because they are computed with different models and from different types of data. Rather, it is changes over time, within each series, that are meaningful. If the Index for mentions increase over some days, then that indicates there is increasing polarization in who the users are talking to or the vocabulary they use.

We combine the individual indices into our single Aggregate INTERPOLAR INDEX by taking the product. We choose this aggregation because it is simple and clear. It reflects the intuition that each component contains useful information, and that an aggregate metric that is proportional to the individual ones

is reasonable. For example, if polarization in text increases while polarization in mentions decreases, both of these changes are important and should be accounted for in our measurement. With the product aggregation, one of the two changes will dominate if it is a proportionally bigger change, or they will cancel out if they are proportionally similar.[13]

In future work, we will consider more complex ways of combining the embeddings and distance measure. In some simple analysis explained in the next section, we find evidence that retweets are strong predictor of party support, while other relations between users are far more mixed. We also plan to retrieve "like" and "reply" relations. We expect that likes will be similar to retweets, but on the other hand, reply relations may again have a mixed effect. A more sophisticated way of combining the different relations may be effective to understand both how much users identify with their own party, and how much they come in conflict with the other party.

4 Findings

This section presents our main findings. We begin with synthetic experiments to examine how our model performs in a controlled setting. We then present experiments on real data.

4.1 Synthetic Experiments

We use a stochastic block model [Holland et al., 1983] to simulate polarized communities. In this graph model, connections are generated randomly according to intra-group and inter-group probabilities. We choose to use two communities (simulating conservatives and liberals or Republicans and Democrats), with a higher intra-group connection probability (i.e. people are more likely to connect to people from the same party). We vary the conditions to examine different aspects of our model.

In all experiments, we generate a network, run the GCN part of our model, then calculate the INTERPOLAR INDEX to estimate polarization. The model used is an unsupervised version, except in the experiment where we explicitly test the effect of adding the supervised component. Unless otherwise noted, we fix the intra-group connection probability at 0.25, the graph size at 500 nodes per community, and repeat the experiments 20 times. The result shown is the mean, with the standard deviation given by the error bars.

First, we consider the basic question of whether the model detects polarization information by varying the inter-group probability between 0.05, 0.10, 0.15, and 0.20. A low inter-group probability produces less interactions with the other group and simulates more polarization, and vice versa. Figure 4 shows that the model behaves as expected: more simulated polarization results in a higher estimated polarization.

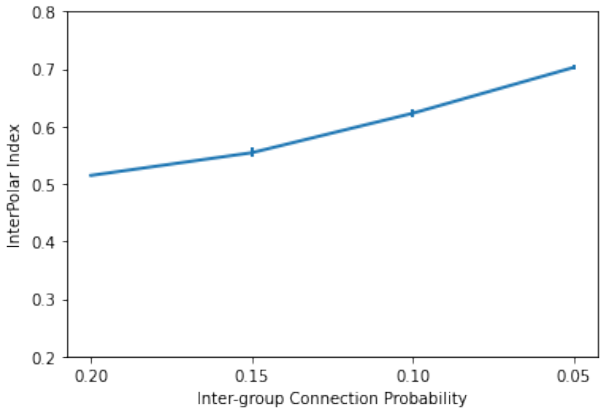

Figure 4: Simulating an increase in polarization by decreasing the interactions between the groups, results in an increase in our estimated polarization.

Next, we keep the same data generating process and compare unsupervised and semi-supervised models. The semi-supervised model is given the correct label for half of the data points from each community, and no label for the rest. We see in Figure 5 that the semi-supervised model can sometimes

[13] Two other simple options are taking the average or the maximum value. However, these can introduce scaling issues, because as noted in the preceding paragraph, the scale itself is hard to interpret between the different relations. For example, if one relation produces INTERPOLAR INDEX which fluctuates within a small range, while another fluctuates in a large range, then the latter can dominate the former in the average. And similarly, if one relation produces INTERPOLAR INDEX which is consistently higher than another, then the latter will be ignored in the maximum. In addition, the maximum can magnify noise in the measurements—the more relations one combines, the more likely the maximum will be just the one with the largest noise. By using the product, we avoid these potential pitfalls.

produce a slightly higher estimated polarization. We hypothesize that this is due to the labels leading the model to further emphasize the differences between the two communities. Nonetheless, the differences identified are small and overall the two versions are quite similar.

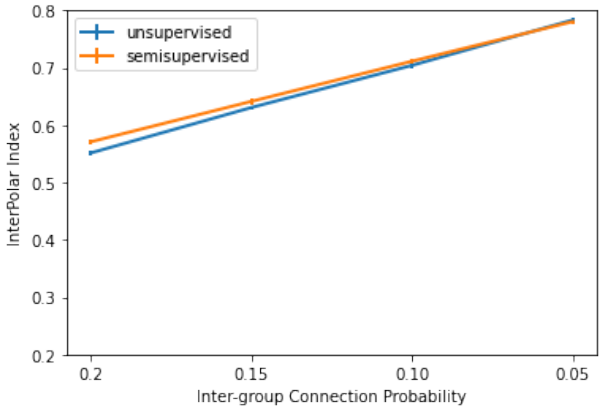

Figure 5: Unsupervised and semi-supervised versions of the model give similar estimates.

Now we consider the effect of imperfect party labels in the estimated polarization. We keep the same data generating process, but corrupt 10, 15, or 20 percent of the labels in the INTERPOLAR INDEX calculation by flipping them to the opposite, incorrect party. This roughly matches our actual accuracy discussed in the real-world experiments. In Figure 6, we first find that the labeling error reduces the level of estimated polarization. This is intuitive, because in the extreme case where users are randomly assigned to a party with equal probability, we expect no polarization at all. But more importantly, while the level is affected, we see that the model still clearly captures changes in polarization. Higher polarization in the simulated data produces higher estimates. Since our goal is to detect and draw conclusions from changes in polarization, rather than measure the static level of polarization, this shows imperfect labels do not pose a significant problem in the data.

We next consider whether the size of the communities impacts the estimates. First, we maintain the balance between the two communities and change both of their sizes equally, from 500 to either 100 or 2500. This simulates a change in overall Twitter activity, such as an important event that would gen-

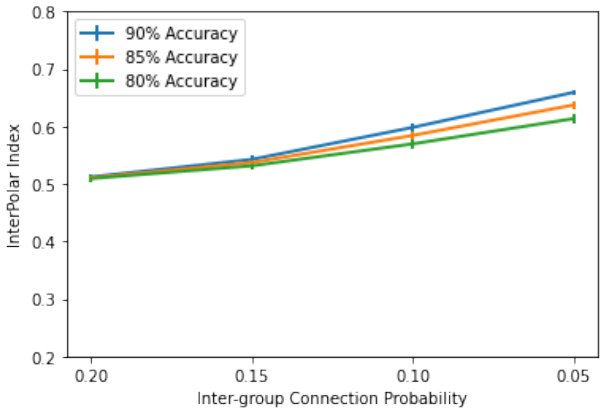

Figure 6: Even with imperfect party predictions, the model detects changes in polarization effectively.

erate discussion from both parties equally. Figure 7 shows this has little impact verified in four different settings. Unsurprisingly the variance is lower when the graph size is large—there is just more data to work with—but it is not too extreme, even at the lowest graph size. Similarly the mean estimates change only marginally; with more simulations the difference likely goes to 0.

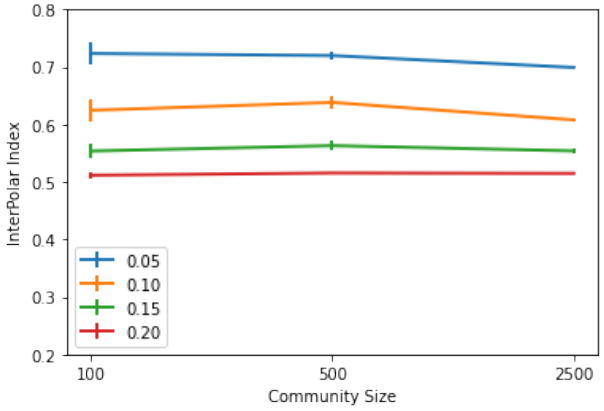

Figure 7: Scaling equal sized communities has little effect on the estimated polarization. Plot shows minimal change in polarization when the size of the communities changes, confirmed in four different settings with varying inter-group connection probabilities.

Instead of keeping the proportional size of each community fixed, we can also vary one community in relation to the other. This simulates, for example, an event that provokes more discussion from one

party than the other. We test this by fixing the size of one community at 500, but varying the size of the other one. Unlike the previous experiment, Figure 8 shows that this has a large impact on the polarization estimate. Depending on the relative sizes of the communities, the estimate can vary by over 100%.

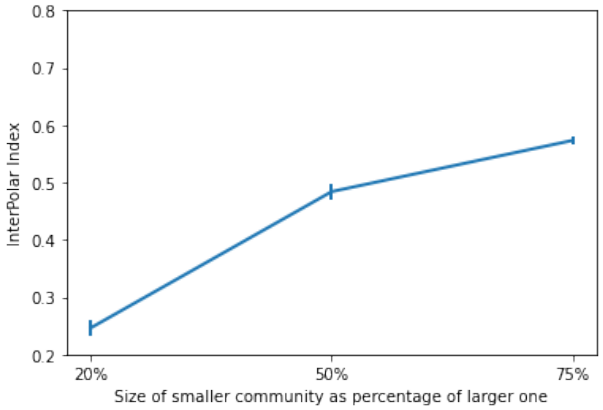

Figure 8: Changing the proportional size of the communities has a large effect on the estimated polarization if not adjusting for the imbalance.

This is very undesirable: with no countermeasure, instead of estimating polarization, we might end up measuring the relative activity level of people from the two parties. To mitigate this, we upsample the smaller community by drawing random nodes ("users") from it, with replacement, until its size matches the larger community. In order to carefully test this upsampling, we expand the number of relative group sizes tested and repeat each test 100 times. Figure 9 shows that upsampling is very effective; the variation in estimated polarization is substantially reduced.

It is not 100% perfect, likely because the upsampling produces some repeated points which show no distance variation in the same community, thereby slightly reducing the estimated polarization. However, in practice, the ratio of community sizes will seldom be as extreme as the 20% (5:1) illustrated here, and the limited variation shown is sufficient to get meaningful estimates. We also tested downsampling, but the variance was slightly higher in this case. Therefore, in all subsequent experiments where we calculate polarization with our model, we use the upsampling strategy.

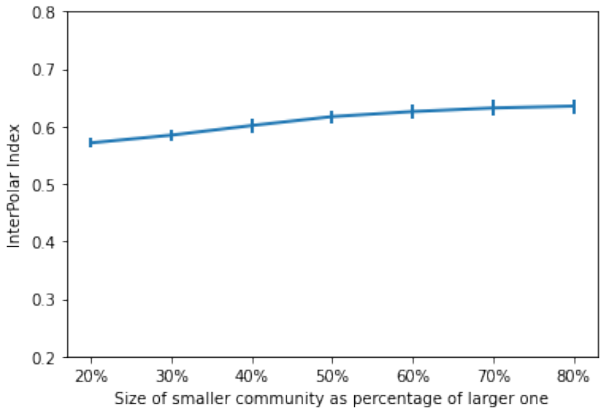

Figure 9: Upsampling fixes incorrect changes in polarization shown in Figure 8 due to community size imbalance.

4.2 Real Data Experiments

We group the real data analysis in three parts:

Question 1: can INTERPOLAR **predict known politician's positions?** Politicians' ideological positions are available based on their voting records. Here we provide a (static) evaluation of our own measures for this specific set of users, politicians, for which we know their DW-NOMINATE scores. We show our embeddings are meaningful by predicting these scores, as well as the politicians' parties.

Question 2: can INTERPOLAR **predict of the partisan affiliations of the mass public?** We evaluate this using voter record matching, our profile labels, and our strong labels.We also perform analysis over time and compare with baselines. We show our model is effective in these predictions.

Question 3: Does polarization **increase around major polarizing events?** We use our embeddings to examine changes over time. We show that changes in polarization correspond to real world events (such as the date of the election or Capitol Hill riot). However, the direction is not always what one might expect.

4.2.1 Politicians

In Table 4 we predict NOMINATE scores and measure the correlation coefficients between these scores and our own measure. We find that there is high variance depending on the train-test split and training of the random forest. Therefore, we report the average and standard deviation of 100 runs with 70-30 random splits. We have a high correlation overall and a weaker but positive correlation within each party.

All	Democrat Only	Republican Only
0.96 ± 0.00	0.27 ± 0.07	0.33 ± 0.08

Table 4: INTERPOLAR embeddings predict ideological scores of members of congress. Table reports the Pearson correlation between DW-NOMINATE scores and our measure.

In Table 5 we show the accuracy of our method in predicting politician's party affiliation. First, we report accuracy using each interaction relation individually (Hashtag, Mention, Retweet, and Quote), and then the accuracy when combining all four. The latter is significantly higher.

Hashtag	Mention	Retweet	Quote	Combined
75.0	81.5	75.2	71.1	91.2

Table 5: INTERPOLAR predicts politician party with high accuracy (%), especially with the combined model user activity embedding.

4.2.2 Mass Public

Turning to the public, in Table 6 we first present some basic information on graphs constructed from Public V1 after filtering for sufficiently active users. The degree numbers are the average number of connections between the respective type of users within the same community ("intra-degree") and the other community ("inter-degree"). Communities are assigned according to the profile labels. The percentage given with the inter-degree is the percent the inter-degree represents of the average total degree (i.e. the average percentage of cross-community connections out of total connections). We see here that retweet has a very low inter-degree relative to the other interaction types.

Next, we return to INTERPOLAR and evaluate the embeddings in two ways. First, we look at the voter registration matched users and compare our predicted party affiliation with the party given by their records. Second, we compare our predicted labels with the strong labels from our domain experts (discussed in 3.1.1).

Results for matched users are shown in Table 7. The distribution of these users is different from the set of profile-labeled users that help our model learn. For example, there are more Democrats here, in contrast to more conservatives in Public V1 as seen in Table 6. But nonetheless we get reasonably accurate predictions.

Table 8 is similar to Table 5, but with the Public V1 dataset. It again shows our method is accurate.

We also performed analyses comparing our model to the TIMME model of Xiao et al. [2020]. On November 3^{rd} alone, our model achieves 91.7% prediction accuracy, while TIMME achieves 93.1% accuracy. Thus, TIMME is slightly more accurate on that particular day. However, TIMME was much slower to run, taking approximately three hours for the single day compared to 30 minutes for our model, on exactly the same hardware. In addition, we have observed cases where our model can run on a much larger amount of data, such as the full Public dataset used in Table 8, while TIMME runs out of memory, again with the same hardware. Thus, our accuracy is quite good and our model is significantly more scalable. This scalability is critical, considering the size of our data and our goal of measuring polarization, not just on one day, but over time. In future work, we will perform additional analyses to compare the two models.

4.2.3 Polarization over time

In this final section, we consider results over time. First, in Figure 10, we present party prediction accuracy. This is similar to Table 8, but while those results were using the entire Public dataset, here we take each day and make a prediction on that day's data alone, with the combined model. Our model achieves a high accuracy, averaging 85%.

We note, however, that there are some days when

Relation	# Cons. Users	# Lib. Users	Cons. Intra-Degree	Cons. Inter-Degree	Lib. Intra-Degree	Lib. Inter-Degree
Retweet	2335	1653	582	29 (4.7%)	294	41 (12.2%)
Mention	3108	2307	1350	760 (36.0%)	863	1024 (54.3%)
Hashtag	1309	998	309	149 (32.5%)	227	195 (46.2%)
Quote	1788	1203	284	81 (22.2%)	164	121 (42.5%)

Table 6: There are significant differences in partisan interactions through the different relations.

New York	Ohio
78.3	78.6

Table 7: INTERPOLAR predicts the voter record affiliation of matched users accurately (%)

Hashtag	Mention	Retweet	Quote	Combined
85.3	88.0	92.6	82.9	93.1

Table 8: INTERPOLAR predicts general public party accurately (%)

the accuracy is lower. This mainly occurs when there is less data available. Because of this, in the subsequent experiments to measure polarization, we use our profile labels when separating users into groups to calculate the polarization indices. This gives a more consistent accuracy and therefore more consistently good polarization measurements. In a future version, we plan to carry forward user embeddings, rather than starting from scratch every day, to get the benefit of the improved accuracy shown in Table 8 without using any post hoc information that the model might not be able to see during real-world usage.

The vertical lines, here and in the subsequent figures, represent November 3^{rd} (election day), December 25^{th}, and January 6^{th} (the day of the Capitol attack).

We next replicate the method of Green et al. [2020] for comparison. This is a random forest model trained on the content of tweets to classify the party of the author. Code was not available but we match their preprocessing routine as closely as possible, with the exception of appending the date of the tweet to the text. Giving the model date information has been found to cause unexpected leakage in other contexts with Twitter data [Pelrine et al., 2021], because the model may learn to rely too heavily on the date instead of more informative information. In this context, for example, if the model learns things like "mostly conservatives tweeted on this day," then the accuracy may no longer reflect performance that one could expect on future days. Therefore, unlike Green et al. [2020], we do not add the date to the text.

We first attempted to train the random forest on our full Public V2 dataset, with over 4 million tweets in the training set, but the training time is large and it was not ready in time for this version of the paper. Instead we present a version trained on 25k tweets randomly sampled from Public V2. This is slightly more than Green et al. [2020] (who trained on $\tilde{2}$1.5k tweets).

Results are presented in Figure 11. Note that contrary to our own analysis, this model is predicting labels per tweet instead of per user. In order to have comparable data points, we predicted the user label, as usual, but then assigned each user's label to all of their tweets. This enables us to exactly match the withheld test set in both types of analyses.

On average our model is 85.3% accurate in this setting, while the random forest model is 72.6% accurate. Our model performs better on almost every day: only 8 out of 122 are worse. Note that our model here is trained on only a single day in the data, while the random forest has access to data from the entire period.

Now that we have shown that our model can deliver meaningful embeddings on a daily basis, we turn to measuring polarization over time. In Figure 12, we show the C index for each type of user embedding from October 2020 to January 2021. Each line can be regarded as a particular kind of polarization. For example, the text line measures the polarization of the language people use, while the mention line represents how polarized their social interactions are.

In Figure 13, we present our overall daily polarization index, which is an aggregation of the indices in the previous figure. This single index is easier to

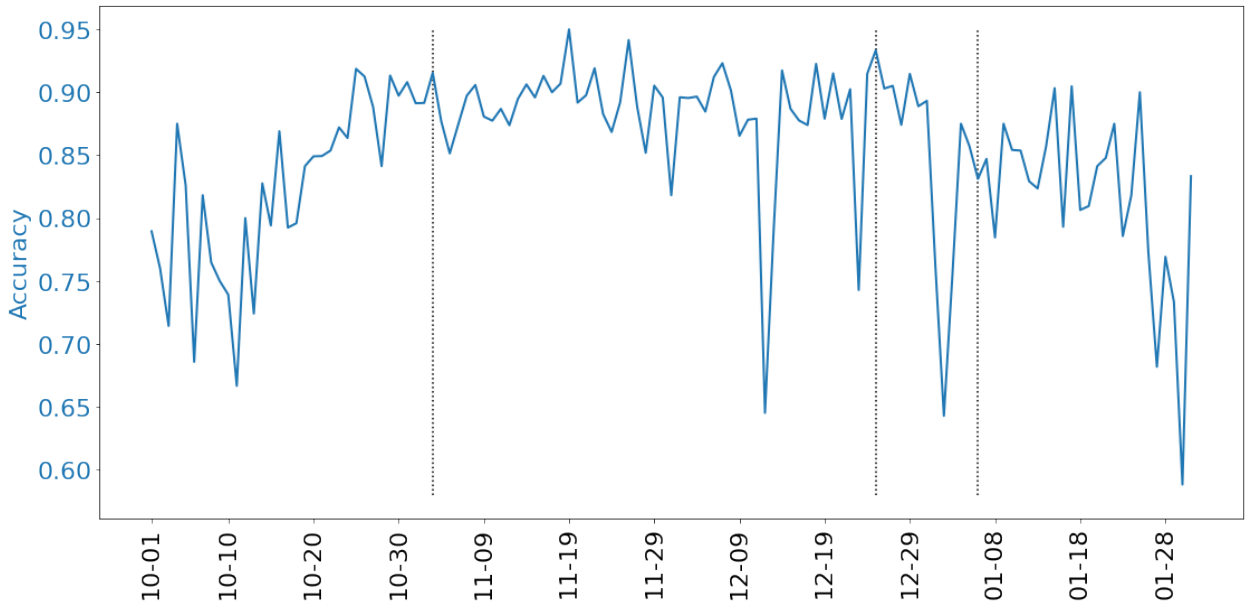

Figure 10: INTERPOLAR predicts party accurately even on individual days

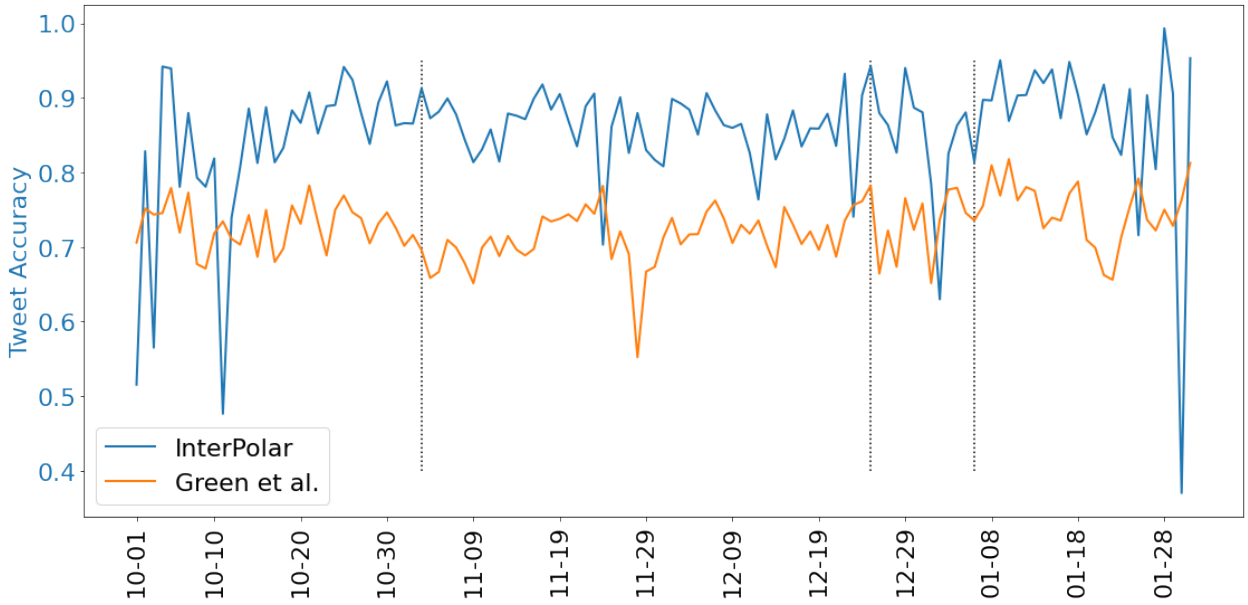

Figure 11: With few exceptions, INTERPOLAR achieves significantly higher accuracy than the baseline.

understand than the previous figure. We discuss how it evolves in the following section.

We next discuss two baselines for potential comparison. First, we consider the approach of Green et al. [2020], which was introduced earlier in this section as a baseline for daily accuracy. The authors use party predictions to measure polarization, arguing that lower accuracy implies that the speech of users from different parties is more similar and therefore less polarized, and vice versa.

However, this is a very strong assumption. Consider a thought experiment where all users from one party write their text (on a hypothetical social media platform) in one color, whereas all users from the other one write in another. But otherwise everything they write is identical. We can classify users with

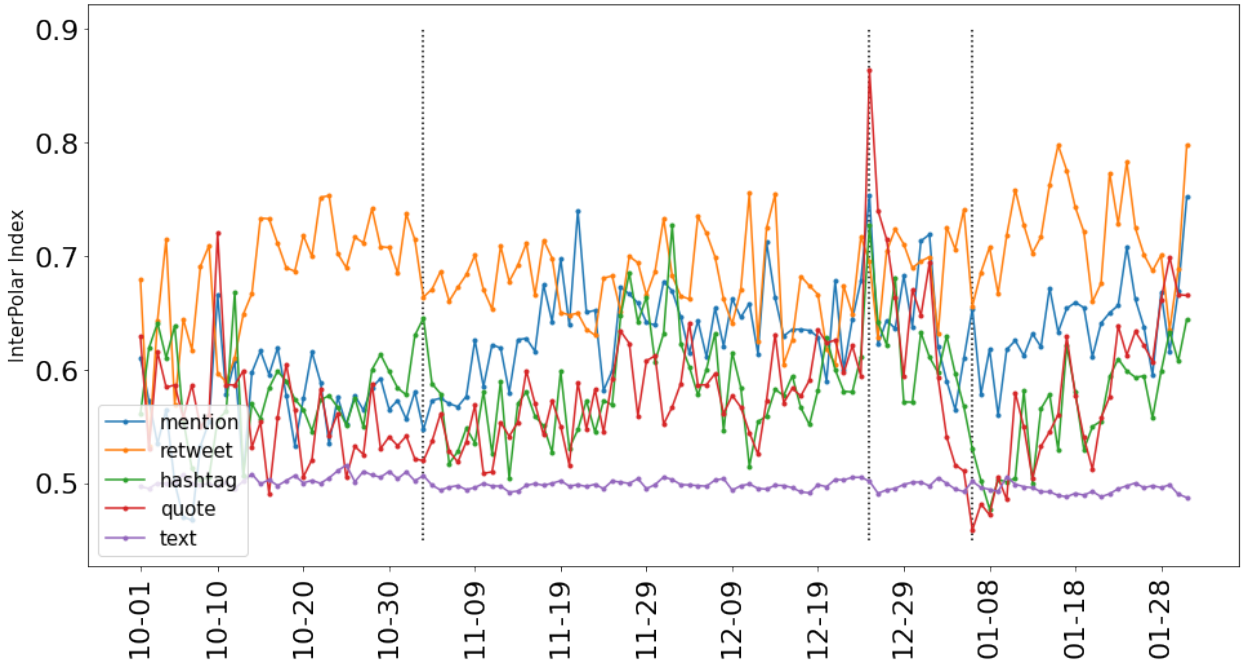

Figure 12: Daily polarization of general public measured using each type of activity in Twitter.

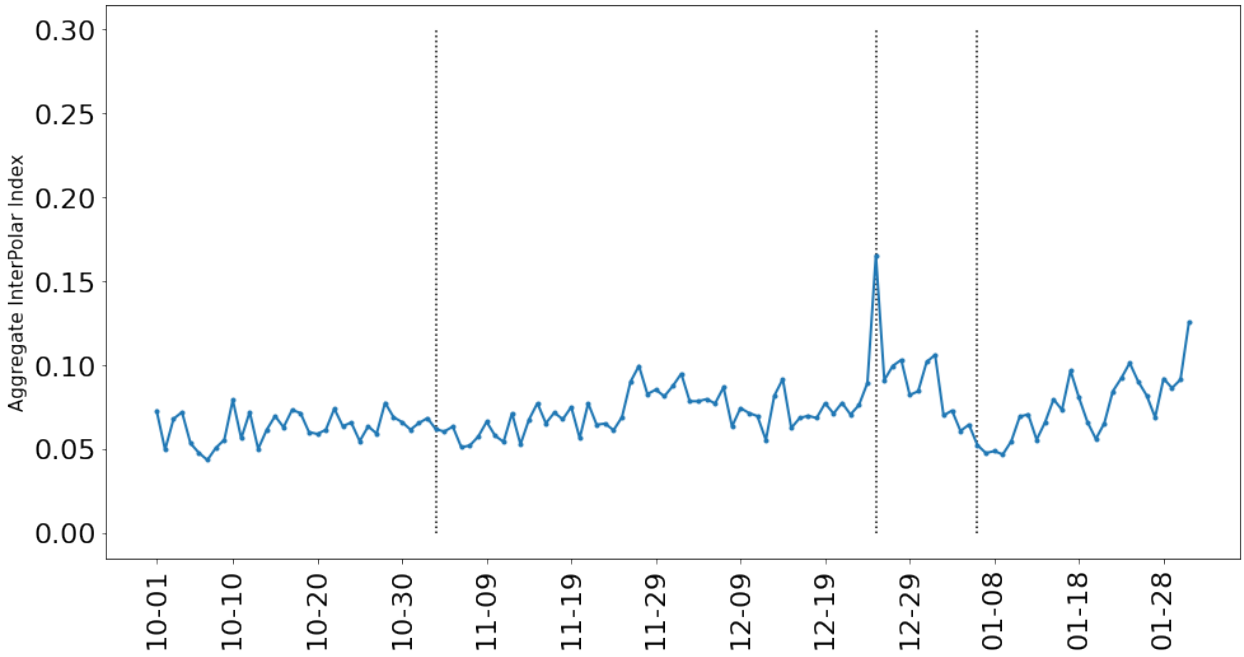

Figure 13: Daily polarization aggregated for all activity types. Polarization increases after the election (1st vertical line), peaks on December 25^{th} (2nd vertical line), then decreases around the Capital Hill attack (3rd vertical line), before increasing again.

100% accuracy based on the color. In this context though, everyone would be saying the same thing, so polarization should be minimal. Thus, it is not clear if higher accuracy is caused by more polar-

ized speech overall, or just a clearer differentiation in some narrow aspect(s) of speech with little overall impact.

In order to test this in practice, we draw on a new dataset we are constructing where users are labeled by intensity of their ideological stance: either moderate or extreme liberals or conservatives. This labeling was done by hand. Experts on our team coded a random sample of users' tweets as moderate or extreme according to whether they contained extreme or conspiracy-related language or hashtags according to a predetermined list.[14] Tweets which contain extreme words were coded as extreme, and the rest as moderate. Users with more extreme than moderate tweets were then classified as extreme, and vice versa moderate. We ultimately aim to use these labels as another form of supervision in our model to directly generate finer-grained ideological estimates for each user. Data collection for that purpose is still in progress, but the 287 users coded so far are sufficient for an experiment here. We first present the basic information on this data in Table 9.

	Users	Tweets
Extreme Liberal	33	14036
Moderate Liberal	119	59431
Moderate Conservative	34	6154
Extreme Conservative	101	19788

Table 9: A new manually-labeled ideology dataset.

Now, if the assumption needed for accuracy to measure polarization holds, we should expect that accuracy will be higher if everyone is extreme. Put differently, accuracy on extreme users should be higher than on moderate users. We test this hypothesis by evaluating the random forest model shown previously on tweets from these users. Results are shown in Table 10.

We see that accuracy is virtually identical between extreme and moderate liberals. It is much worse for conservatives in general, but slightly more accurate for extreme ones. These results are in line with those reported in Green et al. [2020]. They similarly found that liberals were easier to classify.

[14]Extreme language corresponds to keywords like: Obamagate, LockThemAllUp, 25thAmendmentNow, VoterFraud. The complete list of keywords is available upon request.

	Accuracy (%)
Extreme Liberal	85.8
Moderate Liberal	85.8
Moderate Conservative	52.0
Extreme Conservative	56.7

Table 10: Classification accuracy does not correlate well with ideology. An example model based on Green et al. [2020] fails to differentiate extreme and moderate liberals by accuracy.

This provides some evidence against the required assumption from the model used in Green et al. [2020]; their measure of polarization fails to vary among extreme and moderate liberals. In this context, it is impossible to determine if these users are becoming more extreme relative to conservatives. Therefore, it seems that measuring polarization with this approach can be somewhat unreliable. In fact, Green et al. [2020] test a secondary model, which produces a noticeably different polarization estimate (peaking on a different week).

With these issues in mind, we turn to a simpler baseline for measuring polarization: graph modularity [Waugh et al., 2011, Conover et al., 2011]. Rather than a model to find embeddings, this approach looks directly at graph connections within and between communities (in this case provided by our profile labels). Because it only provides a single number rather than an embedding, it cannot be used for more diverse tasks like party predictions, nor use profile labels to improve on these predictions. However, it does have a clear advantage in ease and speed of computation.

Modularity also comes with its own assumptions, which are challenged by Guerra et al. [2013]. They argue that comparing modularity between different datasets is problematic, but we can get around this problem by relying on a single source of data. They also argue that understanding the level of polarization is challenging, but that is not our objective here, since we are interested in changes over time. Thus, while not a perfect measure, we use it here to compare with the results of our own model.

In Figure 14 we present the modularity equivalent of Figure 12.

In Figure 15, we combine the four individual mod-

Figure 14: Modularity baseline with each individual graph.

ularities by multiplication. Note the numbers are quite small due to the product; as before, the important part here is the changes, not the level. Similar to measurements with INTERPOLAR, the modularity index shown here indicates low polarization around January 6^{th}, then a significant increase later in the month.

In the next two figures we plot word clouds for two days of interest, December 25^{th} and January 6^{th}. These can help zoom in and better understand what topics were discussed. First, we show December 25^{th}, where our polarization index peaks. Note this peak is not captured by the modularity index.

Next, we show January 6^{th} (Capitol Hill riot), where measured polarization is low.

We see in Figure 12 that polarization on the 25^{th} occurs particularly in the quote relation, though it is also high in all of the other types of relations (i.e., hashtag, retweet, mention). On January 6^{th}, both quote and hashtag polarizations are particularly low. In order to make sense of this difference, we compare the distribution of these two relations within in each day.

First, Table 11 shows aggregate statistics. The "# Lib." and "# Cons." columns refer to the number of quotes from liberal and conservative users respectively, according to the profile labels. The similar "Shared" columns give the count of the relations from the stated group with a target that the other group also connects to. For example, if both liberal and conservative users quote Person A, then the number of times liberal users quoted Person A will count towards the number in the "Lib. Shared" columns.

We observe significantly more liberal quotes and hashtags on the 6^{th}, compared to the 25^{th}, and the reverse for conservatives, particularly when looking at the quotes. This is likely the result of a combination both of users leaving Twitter (more of whom were conservative) and real-world events. For quotes, we see that the shared percentages are similar between the two days, and slightly lower on the 6^{th}. On the other hand, for hashtag, we find that they are clearly higher on the 6^{th}. Modularity is more complicated but correlated with these numbers, because it compares connections within and between each groups with a degree of randomization in the connections (i.e., configuration model). We see in Figure 14 that the quote and hashtag modularities follow the general pattern in the sharing.

We next examine the most popular quotes and hashtags. Table 12 shows for each day and group the 5 people quoted the most. It also shows the

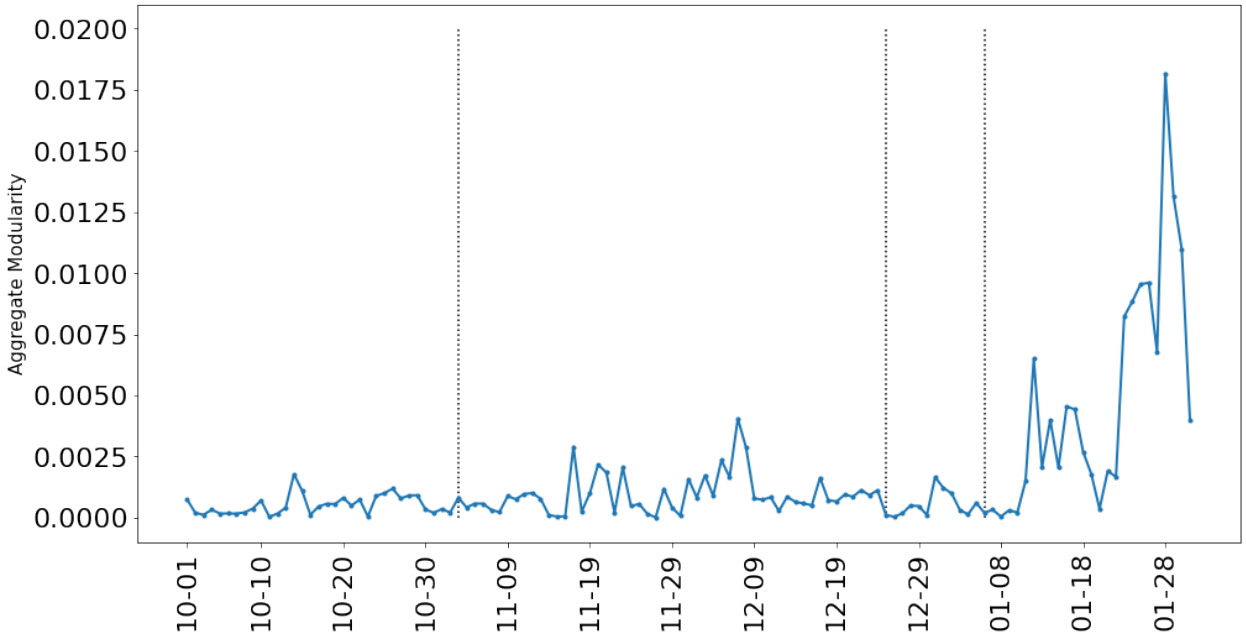

Figure 15: An overall modularity-based baseline. The general trend in January is similar to our model, though more extreme.

Date	Quote				Hashtag			
	# Lib.	# Lib. Shared	# Cons.	# Cons. Shared	# Lib.	# Lib. Shared	# Cons.	# Cons. Shared
Dec. 25	1,582	661 (41.9%)	2,938	1,188 (40.4%)	1,236	436 (35.3%)	1,467	474 (32.3%)
Jan. 6	2,871	1,087 (37.9%)	1,103	457 (41.4%)	4,730	2,057 (43.5%)	1,695	909 (53.6%)

Table 11: The shared percentages for quotes are similar across both days. For hashtags, more are shared on the 6^{th}.

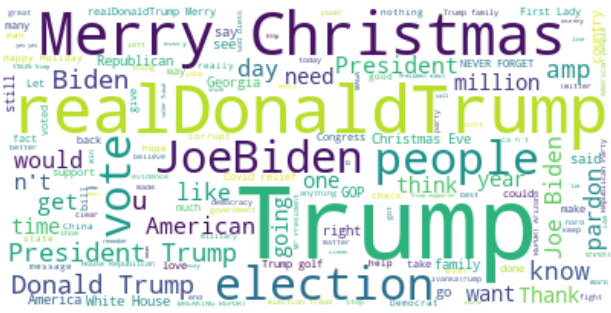

Figure 16: Wordcloud of tweets on December 25^{th}, where measured polarization peaks.

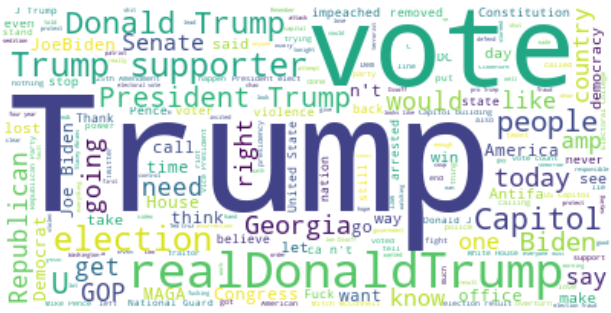

Figure 17: Wordcloud of tweets on January 6^{th}, where measured polarization is low.

amount of times that group quoted each of those people, and the percent of the total quotes from the group each person represents. So for example, the top left entry shows that realDonaldTrump was quoted 102 times, which represents 6.4% of the total quotes by Democrats in our sample that day.

We see a significant difference between the two days, with the counts much more concentrated at the top on December 25^{th}, especially among conservatives. This is a likely cause of the large spike in polarization detected by our model (see Figure 12). This suggests that our model can capture more sophisti-

cated differences in behavior than either the aggregate statistics or modularity modeling approaches.

Finally, table 13 shows the same information but for the hashtag relation. In this case, the differences between the days are less clear. We plan to conduct a more in depth analysis in the future to understand this trend.

5 Summary of Analyses

The overall level of correlation observed between our measures and the NOMINATE scores, shown in Table 4, is very high (.96). While this seems excellent on the surface, we note that a similar correlation can be obtained from linear regression on the politician party alone. However, our method achieves intra-party correlations of .27 (Democratic) and .33 (Republican). This is despite using completely different data sources—either social media activities or NOMINATE scores, which are derived from roll-call votes. While these correlations are not perfect, they show our user activity embeddings are capturing partisan polarization.

We can also accurately predict the party affiliation of Members of Congress, as seen in Table 5, again showing that our user activity embeddings are meaningful. We see a significant improvement from using all four interaction types together, motivating the use of a multimodal method like ours.

Cohen and Ruths [2013] argue that a good performance for the political elite is insufficient to guarantee a meaningful prediction for the general public. In Tables 7 and 8, we see that our method still works effectively when we focus on the mass public by looking at matched primary voting records and strong-labeled users. In addition, despite reducing the size of the data, Figure 10 confirms that performance remains strong when we move to the temporal setting and apply the model on each day individually. We also outperform the baseline shown in Figure 11.

Our results also confirm in Figure 13 that polarization is changing over time on Twitter. First, the results indicate that there was a slight decrease in partisan polarization observed in the days immediately following the November 3^{rd} election. However, we find that the overall level of polarization subsequently increased throughout the rest of November and December, as President Trump and his supporters continued to contest the legitimacy of the vote. It peaked on December 25^{th}. As we saw, there were many interactions involving polarizing users through quotes on Christmas Day. These atypical interactions could thus have played a role in increasing the polarization index in our model. We discussed why INTERPOLAR could detect these interactions while modularity missed them. Surprisingly, we found that polarization actually decreased around the Capital Hill riots on January 6^{th}. We hypothesized that this was mainly due to a significant number of highly polarized and polarizing users leaving Twitter around that time, either by moving to more extreme platforms like Parler, or because their account were suspended. During this time, Twitter suspended over 70,000 accounts,[15] not to mention those who left of their own volition, so this likely had a significant impact on user interactions and polarization.

It is also possible that the conversation became relatively unified in condemning these events, similar to a 'rally round the flag effect,' before diverging again later as competing narratives took root again. However, this would not explain why polarization seems to decrease even before January 6^{th}, so the previous Twitter purge hypothesis could have played a bigger role.

Finally, we saw that later in January, polarization began to increase again. This is likely due to the strong conflict in narratives about the Capitol Hill riot and President Biden's inauguration. In future work, we plan to extend the observation period into February to examine the effects of the second impeachment trial and to look more closely at polarization on the conservative social network Parler.

6 Conclusions and Future Works

This paper introduced a new dynamic framework, INTERPOLAR, to study political polarization on social media platforms. This approach is the first to combine both social interactions and the text content of online messages to estimate a measure of partisan polarization by analyzing more than 365 millions posts on Twitter and Parler during the 2020 presidential campaign. We do this by generating user activ-

[15] https://www.bbc.com/news/technology-55638558

Dec. 25				Jan. 6			
Liberal		Conservative		Liberal		Conservative	
realDonaldTrump	102 (6.4%)	realDonaldTrump	406 (13.8%)	LindseyGrahamSC	68 (2.7%)	ElijahSchaffer	22 (2.0%)
JoeBiden	38 (2.4%)	TimRunsHisMouth	171 (5.8%)	ElijahSchaffer	38 (1.3%)	TheLeoTerrell	21 (1.9%)
donwinslow	33 (2.1%)	marklevinshow	85 (2.9%)	atrupar	35 (1.2%)	JoeBiden	16 (1.4%)
michaelluo	29 (1.8%)	NewDayForNJ	79 (2.7%)	Phil_Lewis_	32 (1.1%)	JackPosobiec	12 (1.1%)
kylegriffin1	25 (1.6%)	JoeBiden	55 (1.9%)	igorbobic	32 (1.1%)	TomiLahren	11 (0.9%)

Table 12: Top 5 most quoted users by each group. On the 25^{th}, there is a clear concentration of quotes in the top 5, especially from conservatives, and quoting polarizing users.

Dec. 25				Jan. 6			
Liberal		Conservative		Liberal		Conservative	
Trump	42 (3.4%)	MAGA	65 (4.4%)	Trump	144 (3.0%)	MAGA	79 (4.7%)
NeverForget	34 (2.8%)	OANN	50 (3.4%)	MAGA	87 (1.8%)	Trump	49 (2.9%)
MerryChristmas	29 (2.3%)	Nashville	31 (2.1%)	25thAmendmentNow	62 (1.3%)	StopTheSteal	42 (2.5%)
EverybodyIsTurningOnTrump	25 (2.0%)	MerryChristmas	30 (2.0%)	ArrestTrump	55 (1.2%)	Georgia	31 (1.8%)
TrumpHatesChristmas	24 (1.9%)	Georgia	27 (1.8%)	Georgia	47 (1.0%)	StopTheSteaI	31 (1.8%)

Table 13: Top 5 most used hashtags by each group.

ity embeddings with a deep language model (roberta-large) and deep graph models (GCNs). In our analyses, we showed that these user activity embeddings effectively capture information on ideology and partisan polarization.

Applying them to measure general public polarization over time, our findings confirm that there was a small decline in partisan polarization after the November 3^{rd} election, followed by a gradual increase in partisan conflicts in the following weeks, with President Donald Trump and his supporters challenging the election results. After a peak on December 25^{th}, polarization decreased temporarily around the January 6^{th} capitol attack, then rose again. In future work, we plan to extend our analysis period into February to examine the second impeachment trial. We also plan to:

- Improve our model using new finer-grained ideology data and more sophisticated methods that preserve user information over time.
- Expand our voter record analysis with more states and stratified analysis.
- Examine other samples of the general public from our large dataset, and construct bootstrap confidence intervals for the measurements.
- Construct a dynamic measure of polarization for Parler users.
- Classify social media users that are independent or non-partisan into a separate analytical category.
- Conduct additional predictive validity tests to check if meaningful events are linked to significant changes in polarization in order to increase our understanding of the peaks and valleys we see in the data.
- Estimate the moderating effects of purging conservative users from the Twitter data.

In the future, we also hope to expand our analytical framework to different social media networks, such as Facebook, Reddit, Instagram and TikTok. Our goal is to develop a model that is scalable to very large datasets and applicable to alternative social media platforms. Finally, we plan to use our measure of polarization in a comparative study to monitor upcoming elections across different countries, including Canada, where we have been collecting data on the most recent election.

References

A. Badawy, E. Ferrara, and K. Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Anal-*

ysis and Mining (ASONAM), pages 258–265. IEEE, 2018. URL https://arxiv.org/pdf/1802.04291.pdf. 4, 5

P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015. 1, 3, 4, 6, 8, 10

P. Barberá, J. Jost, J. Nagler, J. Tucker, and R. Bonneau. Tweeting from left to right. *Psychological Science*, 26:1531 – 1542, 2015. 1, 5

J. Bright. Explaining the emergence of echo chambers on social media: the role of ideology and extremism, 2017. 5

Z. Chen. Mass ideology-based voting model. In *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 871–877, 2015. doi: 10.1109/IAEAC.2015.7428681. 5

R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *Seventh international AAAI conference on weblogs and social media*, 2013. 21

M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE, 2011. 4, 18

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805. 7

D. Diermeier, J.-F. Godbout, B. Yu, and S. Kaufmann. Language and ideology in congress. *British Journal of Political Science*, pages 31–55, 2012. 4

P. DiMaggio, J. Evans, and B. Bryson. Have american's social attitudes become more polarized? *American journal of Sociology*, 102(3):690–755, 1996. 3

M. P. Fiorina and S. J. Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11: 563–588, 2008. 3

F. Gaisbauer, A. Pournaki, S. Banisch, and E. Olbrich. Ideological differences in engagement in public debate on twitter. *Plos one*, 16(3): e0249241, 2021. 5

V. Garimella and I. Weber. A long-term analysis of polarization on twitter. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 528–531. AAAI PRESS, 2017. URL https://www.icwsm.org/2017/. International AAAI Conference on Web and Social Media, ICWSM ; Conference date: 15-05-2017 Through 18-05-2017. 5

A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, et al. A weakly informative default prior distribution for logistic and other regression models. *Annals of applied Statistics*, 2(4):1360–1383, 2008. 1

M. Gentzkow, J. M. Shapiro, and M. Taddy. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340, 2019. 4

N. Gidron, J. Adams, and W. Horne. *American Affective Polarization in Comparative Perspective*. Cambridge University Press, 2020. 3

J. Green, J. Edgerton, D. Naftel, K. Shoub, and S. J. Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. *Science Advances*, 6(28):eabc2717, 2020. 4, 5, 15, 16, 18

N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378, 2019. ISSN 0036-8075. doi: 10.1126/science.aau2706. URL https://science.sciencemag.org/content/363/6425/374. 4

A. Gruzd and J. Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & Internet*, 6(1):28–45, 2014. doi: 10.1002/1944-2866.POI354. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI354. 4

P. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Seventh international AAAI conference on weblogs and social media*, 2013. 18

M. J. Hetherington and T. J. Rudolph. *Why Washington won't work: Polarization, political trust, and the governing crisis*, volume 104. University of Chicago Press, 2015. 3

P. W. Holland, K. B. Laskey, and S. Leinhardt.

Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 11

L. Hubert and J. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080, 1976. 10

S. Iyengar and M. Krupenkin. The strengthening of partisan affect. *Political Psychology*, 39:201–218, 2018. 3

S. Iyengar, G. Sood, and Y. Lelkes. Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431, 2012. 1, 3

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. 9

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2017. 8

A. Klein, L. Tiao, T. Lienart, C. Archambeau, and M. Seeger. Model-based asynchronous hyperparameter and neural architecture search. *arXiv preprint arXiv:2003.10865*, 2020. 10

G. C. Layman, T. M. Carsey, and J. M. Horowitz. Party polarization in american politics: Characteristics, causes, and consequences. *Annu. Rev. Polit. Sci.*, 9:83–110, 2006. 3

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. 7

N. McCarty, K. T. Poole, and H. Rosenthal. *Polarized America: The dance of ideology and unequal riches*. mit Press, 2016. 1, 3

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 10

K. Pelrine, J. Danovitch, and R. Rabbany. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, pages 3432–3441, 2021. 15

K. T. Poole and H. Rosenthal. On party polarization in congress. *Daedalus*, 136(3):104–107, 2007. 3, 6, 10

R. Rabbany, M. Takaffoli, J. Fagnan, O. R. Zäane, and R. J. Campello. Relative validity criteria for community mining algorithms. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 258–265. IEEE, 2012. 10

L. Rheault and C. Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1): 112–133, 2020. doi: 10.1017/pan.2019.26. 1, 4

J. B. Slapin and S.-O. Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008. 4

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 7

L. Vendramin, R. J. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical analysis and data mining: the ASA data science journal*, 3(4):209–235, 2010. 10

A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, and M. A. Porter. Party polarization in congress: A network science approach, 2011. 18

Z. Xiao, W. Song, H. Xu, Z. Ren, and Y. Sun. Timme: Twitter ideology-detection via multitask multi-relational embedding. *arXiv preprint arXiv:2006.01321*, 2020. 9, 14

K.-C. Yang, P.-M. Hui, and F. Menczer. How twitter data sampling biases u.s. voter behavior characterizations, 2020. 4

M. Yang, X. Wen, Y. Lin, and L. Deng. Quantifying content polarization on twitter. In *2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, pages 299–308, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/CIC.2017.00047. URL `https://doi.ieeecomputersociety.org/10.1109/CIC.2017.00047`. 4

M. Yarchi, C. Baden, and N. Kligler-Vilenchik. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, pages 1–42, 2020. 3, 4

S. Yardi and D. Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin*

of science, technology & society, 30(5):316–327, 2010. 5